CAUCHY: Jurnal Matematika Murni dan Aplikasi
Volume 10(1) (2025), Pages 166-179
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Analyzing Household Expenditures with Generalized Random Forests

Eriski Isnanda, Khairil Anwar Notodiputro*, Kusman Sadik

Department of Statistics, IPB University, Bogor, Indonesia

Email: anwarstat@gmail.com

ABSTRACT

This study investigates the performance of Generalized Random Forest (GRF), which has been known to be useful in understanding heterogeneous treatment effects (HTE) and non-linear relationships in high-dimensional data. In this paper the performance of GRF was compared with Random Forest (RF), Generalized Linear Mixed Model (GLMM) as continuation of previous study conducted by Athey (2019). The data utilized in this study is from the National Socioeconomic Survey (SUSENAS) to predict household per capita expenditure in West Java, Indonesia. The models are evaluated based on their ability to handle outliers using Winsorization. The results show that RF performed the best, yielding the smallest MSE values, followed by GRF with reasonably good performance, and GLMM with the highest MSE, indicating its limitations in handling non-linear data patterns. These findings indicate that RF is the most accurate method for modeling per capita expenditure in West Java, with recommendations for further research to develop hybrid methods or use more specific random effects in GLMM.

Keywords: generalized linear mixed model; generalized random forest; household per capita expenditure; random forest; winsorization

INTRODUCTION

In the rapidly developing digital era, the ability to analyze and predict data has become an increasingly important skill. Machine learning (ML) and statistical methods play a crucial role, especially in analyzing data and building accurate predictive models [1]. Machine learning allows computers to learn from data and make predictions or decisions without explicit programming, using methods such as supervised learning and unsupervised learning [2]. On the other hand, statistical methods provide a theoretical foundation to understand and model data, measure uncertainty, test hypotheses, and identify relationships among variables [3]. Statistical methods also help test model validity and ensure that predictions generated by machine learning are not only accurate but also accountable [4]. Both approaches complement each other in addressing the challenges of increasingly complex and dynamic data analysis across various fields.

One of the popular methods today is the Generalized Random Forest (GRF). The Generalized Random Forest (GRF) method is a machine learning technique that combines

Submitted: November 29, 2024 Reviewed: December 13, 2024 Accepted: March 03, 2025
DOI: http://dx.doi.org/10.18860/ca.v10i1.30104

principles from Random Forests (RF) with more in-depth statistical techniques. The RF method is an ensemble-based machine learning algorithm that uses multiple decision trees to make predictions. The GRF method develops this idea by offering greater flexibility and allowing for the handling of various statistical problems, such as evaluating heterogeneous treatment effects and causal analysis [5].

The advantage of the GRF method lies in its ability to handle heterogeneous treatment effects, which refers to efforts to measure and understand how the impact of a treatment, policy, or intervention varies among individuals or subpopulations within a population [6]. For example, Goldman used the GRF method to evaluate the impact of types of care (hospitalized vs. home care) on suicide risk. The study found that hospitalized patients had a higher suicide risk than those discharged. This heterogeneous treatment effect helps provide a better understanding of variations among individuals or groups in data [7]. Other studies that have also utilized GRF include [8], [9], and [10].

Research on the performance of GRF using Indonesian data has not been conducted. Therefore, it is interesting to study the performance of GRF based on data from Indonesia. In this case, the performance evaluation will be conducted by comparing the GRF method with other well-known methods, particularly RF and Generalized Linear Mixed Model (GLMM). These two methods are chosen as comparatives because RF forms the basis of GRF, while GLMM is a popular method in statistics.

In Indonesia, SUSENAS data collected by BPS every March and September covers samples from various census blocks and contains important information such as per capita expenditure, which can be analyzed to understand the economic welfare of society. The GRF method can be used to explore the complex relationships between per capita expenditure and factors such as household characteristics, geographic location, education, and employment sector. RF and GLMM can also analyze per capita expenditure, but RF is less effective in causal inference, while GLMM has limitations in handling non-linear relationships and strict distribution assumptions. However, GLMM excels in accommodating random effects in hierarchical data structures.

The RF method builds several decision trees to generate stronger and more reliable predictions. The strength of RF lies in its ability to handle complex interactions between variables without needing specific assumptions about the form of the relationship between explanatory variables and the target [11]. Studies such as [12], [13], and [14] have shown that RF improves accuracy in data predictions. Additionally, GLMM combines fixed and random effects, enabling analysis of variability between groups in data [15]. This method is highly useful for longitudinal data analysis or data with unobserved effects. Some studies using this method include [16] and [17].

This study compares the GRF, RF, and GLMM methods using per capita household expenditure data from the March 2021 SUSENAS survey in West Java. Per capita expenditure has a right-skewed distribution, with most households having low expenditures and a small proportion with very high expenditures, leading to outliers. Previous research, such as Belinda's study, shows that outliers may reflect real phenomena rather than errors and should not be discarded but handled using Winsorization. Winsorization reduces the impact of extreme values without removing data, preserving distribution characteristics. In this study, Winsorization was applied based on the Interquartile Range (IQR), adjusting values outside the first and third quartiles, enhancing estimate stability and accuracy.

Based on this, this study focuses on comparing the performance of the GRF, RF, and GLMM methods in modeling household per capita expenditure in West Java. The selection of West Java in this study is just an example, and similar analysis can be applied to other regions with different characteristics. This study uses data that includes various explanatory variables to assess the ability of each method to generate accurate predictions, while also considering the strengths and weaknesses of each model.

METHODS

Data

The data used in this study comes from the 2021 National Socioeconomic Survey (SUSENAS) conducted by the Central Statistics Agency (BPS) of West Java Province. The data includes a response variable (Y), which represents household per capita expenditure for one month, available at the household member level. This per capita expenditure is calculated by dividing the total household expenditure for one month by the number of household members. Per capita expenditure data is crucial for analyzing the economic welfare of the population, as it reflects the household's consumption capacity. Additionally, there are predictor variables (X) assumed to represent household per capita expenditure, with a total of 22 explanatory variables. All variables used in this study, including the response and predictor variables, are presented in Table 1.

Table 1. Definition and Classification of Variables

Variable	Variable Description	Type of variable	Scale
EXPENDITURE	Household per capita expenditure in a month	Response	Ratio
REGION	Type of region	Predictor	Nominal
GENDER	Gender of household head	Predictor	Nominal
AGE	Age of household head	Predictor	Ratio
EDU	Highest education of head of household	Predictor	Ordinal
SAVING	Percentage of family members having saving account	Predictor	Ratio
ILLITERATE	Percentage of family members illiterate	Predictor	Ratio
FOOD	Household food insecurity status	Predictor	Nominal
PLACE	Ownership Status of Place	Predictor	Nominal
HOUSE	House Size	Predictor	Ratio
OTPLACE	Have another place to stay	Predictor	Nominal
ROOF	Roof types	Predictor	Nominal
WALL	Wall types	Predictor	Nominal
FLOOR	Floor types	Predictor	Nominal
DEFECATION	Defecation facilities	Predictor	Nominal
WATER	Drinking Water Source	Predictor	Nominal
ELECTRICITY	Electricity	Predictor	Nominal
CREDIT	Household receives one type of credit	Predictor	Nominal
LAND	Ownership of land	Predictor	Nominal
INCOME	Main income from the transferee	Predictor	Nominal
KKS	Have card "Keluarga Sejahtera"	Predictor	Nominal
AID	Receive social aid	Predictor	Nominal
MICRO	Have a micro enterprise	Predictor	Nominal

This study focuses on several regencies/cities in West Java with both high and low Gross Regional Domestic Product (GRDP), which are expected to represent per capita expenditure in the province. The regencies/cities with high GRDP include Bekasi regency, Bogor regency, and Bandung city, while those with low GRDP include Sukabumi city, Banjar city, and Kuningan regency. The sample size includes 1,265 households in Bogor Regency, 1,147 households in Bekasi Regency, 1,113 households in Bandung City, 638 households in Sukabumi City, 567 households in Banjar City, and 831 households in Kuningan Regency.

Random Forest

Random Forest is an ensemble-based decision tree method that combines predictions from multiple trees to produce a more accurate model that is resistant to overfitting. In the context of regression, this method is particularly effective in predicting continuous outcomes by averaging the predictions from individual trees, which helps improve accuracy and stability [11].

The process of building a model using the RF algorithm involves several steps. Let the number of trees to be formed be denoted as B. For each tree $(b = 1, \dots, B)$, a bootstrap sample of size n is randomly drawn with replacement from the training data (D) to obtain D_b^*. For each tree, the following steps are performed at the t-th node with following conditions : (i) randomly select $m \approx \frac{1}{2}\sqrt{p}$, $m \approx \sqrt{p}$ or $m \approx 2\sqrt{p}$ predictor variables, (ii) determine the best splitting criterion which is the split point that minimizes the Mean Squared Error (MSE), (iii) split the data at node t based on the splitting criterion in step ii, repeat steps (i) to (iii) until the stopping criteria are met to obtain the estimated result for a single tree. The predictions from each tree $\hat{Y}_b(X)$ for data $X = \left(X_1, X_2, \dots, X_p\right)$ are then averaged to obtain the final prediction from all the trees in the Random Forest:

$$\hat{Y}(X) = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_b(X) \tag{1}$$

Generalized Linear Mixed Model (GLMM)

The Generalized Linear Mixed Model (GLMM) is an extension of the Generalized Linear Model (GLM) that combines both fixed effects and random effects [15]. Fixed effects $(X\beta)$ represent the global relationship that applies to the entire population, while random effects (Zb) capture the variation between groups or clusters that cannot be explained solely by the fixed effects. GLMM is used for repeated measures data, hierarchical data, or data with a grouped structure. The GLMM model is defined as follows:

$$g(\mu) = X\beta + Zb \tag{2}$$

The link function $g(\mu)$ connects the mean response (μ) with the predictors (X, Z). Fixed effects $(X\beta)$, a linear combination of fixed predictors and their coefficients (β). Random effects (Zb), where b is a random variable following a normal distribution: $b \sim N(0, \sigma^2)$. There are several steps in building a GLMM model:

1. Identify the response variable (Y) and predictor variables (X)
2. Classify the predictors into fixed effects and random effects
3. The probability distribution for the response variable (Y) uses a Gaussian distribution because per capita expenditure is continuous data
4. The link function used is the identity $(g(\mu) = \mu)$, so the model can be written as :

$$Y_{ij} = \beta_0 + \sum_{k=1}^{p}\beta_k X_{ijk} + \sum_{l=1}^{q} b_l Z_{ijl} + \varepsilon_{ij} \tag{3}$$

where the response value Y_{ij} for the j-th observation in the i-th group, the k-th fixed predictor $\left(X_{ijk}\right)$, the l-th random predictor $\left(Z_{ijl}\right)$, random effects(b_l), $b_l \sim N(0,\sigma^2)$, and residual error$\left(\varepsilon_{ij}\right)$, $\varepsilon_{ij} \sim N(0,\sigma^2)$.

5. Parameter Estimation is performed using the Maximum Likelihood Estimation (MLE) or Restricted Maximum Likelihood (REML) method. The steps are as follows:

- Likelihood function
  Define the likelihood function based on the Gaussian distribution for Y :

$$L(\beta,b) = \prod\nolimits_{i=1}^{N} \prod\nolimits_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(Y_{ij}-\mu_{ij}\right)^2}{2\sigma^2}\right) \tag{4}$$

  where:

$$\mu_{ij} = \beta_0 + \sum\nolimits_{k=1}^{p} \beta_k X_{ijk} + \sum\nolimits_{l=1}^{q} b_l Z_{ijl} \tag{5}$$

- Log-likelihood function
  Convert the likelihood function into the log-likelihood function to simplify optimization:

$$\ell(\beta,b) = -\frac{1}{2}\sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{n_i} \left[\log(2\pi\sigma^2) + \frac{\left(Y_{ij}-\mu_{ij}\right)^2}{\sigma^2}\right] \tag{6}$$

- Iterative optimization
  Estimate the parameters β (fixed effects) dan b (random effects) iteratively using the Newton-Raphson or Expectation-Maximization (EM) algorithm to find the maximum log-likelihood. Then, calculate the random effects b based on the posterior distribution:

$$b{\sim}N(0,\sigma_b^2) \tag{7}$$

6. Model validation by performing residual analysis using standardized residuals. Check the model quality using Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), and evaluate prediction accuracy using Mean Squared Error (MSE).
7. Predict the values of Y_{ij} using the estimated parameters:

$$\hat{Y}_{ij} = \hat{\beta}_0 + \sum\nolimits_{k=1}^{p} \hat{\beta}_k X_{ijk} + \sum\nolimits_{l=1}^{q} \hat{b}_l Z_{ijl} \tag{8}$$

Generalized Random Forest (GRF)

Generalized Random Forest (GRF) is an extension of Random Forest (RF) designed to expand the application of decision trees into the context of statistical inference. GRF enables in-depth analyses such as estimating causal effects (Conditional Average Treatment Effect, CATE), quantile regression, and other nonparametric models. This technology focuses on local estimation using adaptive weighting, allowing for more relevant and accurate conclusions for each individual in the dataset (Athey et al., 2019). Fundamentally, GRF differs from RF in its analytical objectives and inferential approach. RF is designed to generate accurate predictions by aggregating results from multiple decision trees to create a stable global estimate. In contrast, GRF focuses on estimating local parameters, such as treatment effects or quantile regression.

In GRF, decision trees are constructed using a splitting method that considers the needs of local estimation, giving greater weight to data relevant to the specific inference. The technical approach of GRF also differs from RF in terms of weighting and tree structure. RF splits the data based on criteria such as Gini impurity reduction or minimum squared deviation without considering the requirements of local estimation. Meanwhile, GRF is designed to ensure that data relevant to local parameter estimation receive higher weights. In other words, GRF is more than just a predictive tool. It is a deep inferential method for answering data-driven questions. The GRF method is used to model more complex relationships, particularly with data involving heterogeneity and interactions between features that are difficult to handle using standard linear regression models. GRF adopts the principles of Random Forest but introduces more complex parameter estimation and allows for the handling of continuous variables.

Several steps are involved in building a model with the GRF algorithm. Let the number of trees to be formed be B. For each tree $(b = 1, \dots, B)$, take a random sample (bootstrap sampling) of size n with replacement from the training data (D) to obtain D_b^*. For each tree, perform partitioning at node t with the following condition : (i) randomly select $m \approx \frac{1}{2}\sqrt{p}, m \approx \sqrt{p}$ or $m \approx 2\sqrt{p}$ predictor variables, (ii) determine the best splitting criterion which is the split point that minimizes the Mean Squared Error (MSE), (iii) split the data at node t based on the splitting criterion from step ii, (iv) repeat steps i through iii until the stopping criterion is met to obtain the estimates from one tree, (v) perform local linear model estimation to predict the response values based on features. A linear model is used to handle heterogeneity issues in predictions, which are common in continuous data :

$$Y_i \approx X_i\beta_b + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2) \tag{9}$$

where β_b is the coefficient estimated for leaf b and ε_i is the residual, (vi) The prediction for new data X^*, is calculated as $\hat{Y}_b{}^* = X^*\hat{\beta}_b$, (vii) The final prediction is obtained by averaging the predictions from all trees in the Generalized Random Forest:

$$\hat{Y}(X) = \frac{1}{B}\sum_{b=1}^{B}\hat{Y}_b(X) \tag{10}$$

Detection Outlier

An outlier is data that is far outside the range of values considered normal or typical in a dataset. This data is very different from other data points, either because of extreme values or because of its unusual nature within the context of the data being analyzed. Outliers may occur due to measurement errors, data recording mistakes, or truly unique phenomena that warrant further analysis [19] and [20]. Outliers can affect statistical analysis results because they can distort calculations of mean, standard deviation, and prediction models. Therefore, identifying and handling outliers is crucial to ensure the accuracy and reliability of data analysis models [21].

Outlier detection using standardized residuals is a commonly used method in regression analysis and other statistical models to identify values that are far different from the model's predictions. Residuals are the differences between observed values and the predicted values from the model, whereas standardized residuals are residuals that have been normalized so they can be compared across data with different scales. By using standardized residuals, we can assess how much the data deviates from the model and determine if it can be considered an outlier [22].

The process of calculating standardized residuals involves dividing the residuals by the estimated standard deviation of the residuals for each data point. The formula used is :

$$Standardized\ Residual = \frac{e_i}{\hat{\sigma}(e_i)} \tag{11}$$

where e_i is the residual for the i-th data point (the difference between the observed value and the prediction), $\hat{\sigma}(e_i)$ is the estimated standard deviation of the residuals. Standardized residual with values greater than 3 or less than -3 are often considered outliers. This means the data is more than three standard deviations away from the predicted value, indicating a significant deviation from the expected pattern [23].

By addressing outliers flagged by standardized residuals, analysts can reduce bias and improve predictive performance. This technique is highly useful in regression analysis and other statistical models involving variables with different scales. Moreover, this method is more sensitive to extreme values, which is crucial for improving model accuracy and reducing the influence of outliers on the analysis results [22] [23].

Winsorization

Winsorization is a statistical technique used to reduce the influence of outliers by replacing them with values closer to other data points, using thresholds determined based on the data distribution, such as the Interquartile Range (IQR). The main advantage of Winsorization over other methods like outlier removal or data transformation is its ability to retain all existing data, reducing the risk of losing important information. However, its limitation lies in the potential reduction of data variability, which may affect the analysis results, especially if the data is highly skewed. Nonetheless, Winsorization remains an efficient and practical choice in many cases, as it does not require structural changes to the data and still maintains the stability of estimates.

The first step in the winsorization process using IQR is to calculate the IQR itself, which is the difference between the first quartile (Q1) and the third quartile (Q3). The IQR measures the middle 50% of the data, which is not influenced by extreme values. After that, two thresholds are calculated to detect outliers, namely the lower and upper bounds. The lower bound is determined as $Q_1 - 1.5 \times \text{IQR}$, while the upper bound is determined as $Q_3 + 1.5 \times \text{IQR}$. Values outside these bounds are considered outliers.

The winsorization process replaces the values detected as outliers. If a value is smaller than the lower bound, it is replaced with the lower bound. Conversely, if a value is larger than the upper bound, it is replaced with the upper bound. For example, if a dataset has extreme values like 100 and 120, which are much higher than the calculated upper bound, these values will be replaced with the corresponding upper bound. In this way, the data becomes more controlled and not distorted by outliers, even though the dataset size is preserved [24].

Data Analysis Procedure

This study will estimate the per capita expenditure variable for households in regencies/cities in West Java. The study uses R version 4.4.1 software with various available R packages. The data analysis procedure is as follows:

1. Pre-processing and data exploration.
2. Model fitting using Random Forest, Generalized Random Forest, and Generalized Random Forest methods. Subsequently, outlier detection will be performed using standardized residuals.
3. Handling outliers using the Winsorization method.
4. Splitting the data into two parts: 70% for training data and 30% for testing data
5. Fitting models on the training data using Random Forest, Generalized Linear Mixed Model, and Generalized Random Forest.

6. Predicting on the test data.
7. Calculating the prediction accuracy of the models based on defined.
8. Evaluating the performance of the models from the three models.
9. Interpreting the results.

RESULTS AND DISCUSSION

Data Exploration

The distribution of per capita household expenditure data as a whole (for Bogor Regency, Bandung City, Bekasi Regency, Kuningan Regency, Banjar City, and Sukabumi City) can be seen in Figure 1. From the histogram in Figure 1, it can be observed that the per capita household expenditure data is skewed to the right. This is due to the presence of several very high or extreme values, which are most likely outliers.

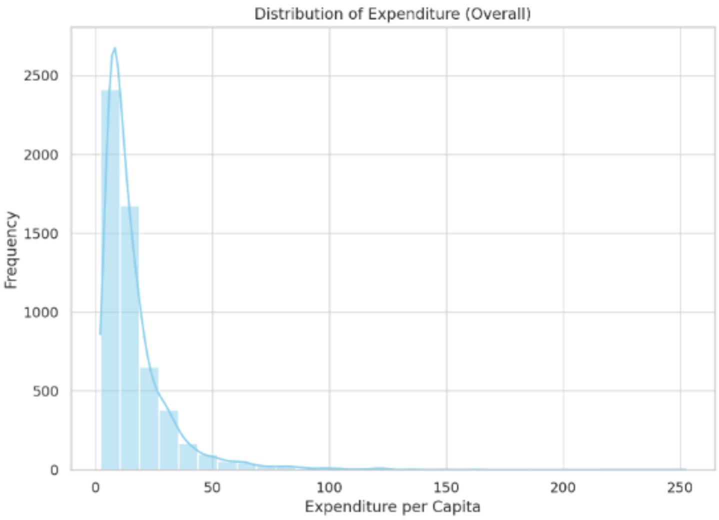

Figure 1. Distribution of household per capita expenditure overall (x 100,000 IDR) in the form of histogram and density plot

Before performing outlier detection, a test was conducted to analyze the correlation between the continuous variables [25]. The analysis results shown in Figure 2 indicate that the four continuous variables have very weak correlations, with correlation values ranging from -0.03 to 0.2. Therefore, it can be concluded that there is no significant multicollinearity among these continuous variables.

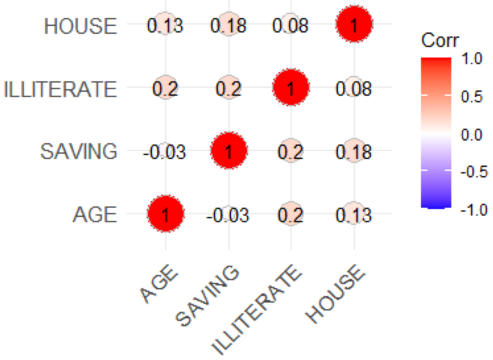

Figure 2. Correlation value between continuous variables

Outlier detection was performed by directly modeling the response variable of household per capita expenditure using 22 predictor variables. The outlier detection was done by observing the standardized residual values [23] generated from the modeling. For this purpose, three different methods were applied: Random Forest (RF), Generalized Linear Mixed Model (GLMM), and Generalized Random Forest (GRF). Each modeling approach was applied to data from each district/city, enabling the analysis of whether there are any outliers in household per capita expenditure in each of these areas. With this approach, it is hoped to gain a deeper understanding of the distribution of household per capita expenditure and identify outliers that may influence the analysis results.

Table 2. Winsorization

GRDP Category	MSE before *Winsorization*			MSE after *Winsorization*		
	GRF	RF	GLMM	GRF	RF	GLMM
High GRDP	147.86	147.01	164.01	82.54	80.62	93.46
Low GRDP	90.34	87.20	95.84	47.23	44.77	53.84

Table 2 above shows a comparison of the average Mean Squared Error (MSE) values from three model methods, namely Generalized Random Forest (GRF), Random Forest (RF), and Generalized Linear Mixed Model (GLMM), before and after Winsorization. From the ten repetitions that resulted in average MSE values, it can be said that Winsorization successfully reduced the MSE values in all regions and for all methods, indicating that Winsorization is effective in mitigating the influence of outlier values on household per capita expenditure data. As seen in the High GRDP Category, there was a noticeable decrease in the MSE values after Winsorization. With the GRF method, the MSE before Winsorization was 147.86, which decreased to 82.54 after Winsorization. The RF method showed similar results, with an MSE of 147.01 before Winsorization, and 80.62 after Winsorization. Meanwhile, the GLMM method recorded a reduction in MSE from 164.01 before Winsorization to 93.46 after Winsorization. It can be seen that Random Forest produced the smallest MSE values for both High and Low GRDP. Further analysis is needed to evaluate the comparison of the three methods to determine if there are significant differences in their performance.

Model Comparison

In order to draw valid conclusions regarding the performance differences between the GRF, RF, and GLMM methods, hypothesis testing is required. Through this hypothesis testing, it will also be possible to determine whether there is an interaction between these three methods and GRDP in producing the MSE of the methods.

Two-way Analysis of Variance (Two-way ANOVA) is a statistical technique used to examine the effect of two independent factors on a dependent variable. In this case, the independent factors analyzed are the GRDP levels (high and low) and three specific methods (GRF, RF, and GLMM), while the dependent variable is the MSE value. This approach allows for testing the main effects of each factor, as well as the interaction between them. The results of the ANOVA test are shown in Table 3.

Table 3. ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	F-Statistic	P-Value
Method	1345	2	10681	< 0.001
GRDP	20454.54	1	324937	< 0.001

Method:GRDP	55.14	2	437.97	< 0.001
Residual	3	54	-	-

Table 3 presents the results of a two-way ANOVA evaluating the effect of modeling methods (GRF, RF, GLMM) and GRDP categories (High, Low) on Mean Squared Error (MSE). The GRDP factor has the greatest influence, with a sum of squares of 20454.54 and an F-Statistic of 324937, while the Method factor is also significant, with a sum of squares of 1344.74 and an F-Statistic of 10681.17 (P-Value < 0.001). The interaction between Method and GRDP is also significant (F-Statistic 437.97, P-Value < 0.001), indicating that the effect of the modeling method on MSE depends on the GRDP category. Additional analysis for the low GRDP group shows a significant difference in MSE among the GRF, RF, and GLMM methods (F-Statistic 4100, P-Value < 0.001), suggesting that these three methods perform differently in a statistically significant manner. As a follow-up to the ANOVA test, a post-hoc Tukey test was conducted to identify specific method pairs with significant differences. The results indicate that in both the Low and High GRDP groups, the best-performing method is RF, as it produces the lowest MSE value, followed by GRF as the next best method, while GLMM demonstrates the lowest performance.

Model Prediction

At this stage, estimates were made for the first 10 observations from each regency/city using the Generalized Random Forest (GRF), Random Forest (RF), and Generalized Linear Mixed Model (GLMM) models. The purpose of these estimates is to understand how the three methods perform on the same data and to evaluate the differences in the results obtained. The prediction results are shown in Table 4 and Table 5.

Tables 4 and 5 show the 10 observations comparing the actual monthly household per capita expenditure values (in Rp 100,000) with the predictions generated by three methods: Generalized Random Forest (GRF), Random Forest (RF), and Generalized Linear Mixed Model (GLMM) for High GRDP and Low GRDP. To determine which method produces the best estimates, the comparison is made based on the smallest sum of squared errors.

The calculation of the sum of squared errors (SSE) shows that the RF method performs the best with the smallest SSE value, indicating the lowest prediction error compared to the other methods. Meanwhile, the GRF method has a slightly higher SSE than RF but lower than GLMM, which indicates that although its performance is better than GLMM, it is still not as optimal as RF. The GLMM method produces the largest SSE, indicating that this method has the highest prediction error and lower accuracy compared to the other two methods. These results suggest that RF provides the best predictions, followed by GRF, and GLMM as the method with the lowest performance. To estimate household per capita expenditure based on High and Low GRDP per capita, Mean Squared Error (MSE) from the predictions was used. The MSE calculation results show that the MSE for Low GRDP per capita, for all methods (GRF, RF, and GLMM), is smaller compared to High GRDP per capita. This indicates that the prediction error for the Low GRDP category is smaller compared to the High GRDP category.

Table 4. High GRDP

No.	**Regencies/Cities**	**The average monthly household per capita expenditure (in IDR 100,000).**			
		Actual	GRF Prediction	RF Prediction	GLMM Prediction
High GRDP					
1	Bandung City	4.4	11.42	10.85	7.73
2	Bandung City	9.82	11.79	10.03	10.24
3	Bandung City	81.01	46.69	63.28	49.78
4	Bandung City	13.45	14.3	13.88	20.04
5	Bandung City	41.74	24.11	32.39	26.21
6	Bandung City	6.19	8.95	7.69	11.23
7	Bandung City	8.98	11.89	10.77	11.56
8	Bandung City	8.6	17.47	16.52	18.71
9	Bandung City	26.03	25.12	26.61	25.71
10	Bandung City	20.94	27.81	27.24	30.05
1	Bekasi Regency	17.71	16.59	16.8	20.13
2	Bekasi Regency	15.01	10.64	12.27	7.99
3	Bekasi Regency	5.93	11.58	9.97	5.36
4	Bekasi Regency	11.45	12.52	12.36	13.11
5	Bekasi Regency	8.57	12.82	11.13	18.66
6	Bekasi Regency	16.03	16.14	15.9	22.02
7	Bekasi Regency	8.53	15.36	14.96	21.39
8	Bekasi Regency	7.45	9.08	7.91	12.63
9	Bekasi Regency	35.62	15.79	21.93	18.07
10	Bekasi Regency	6.06	9.04	7.64	6.28
1	Bogor Regency	3.91	11.69	8.36	12.07
2	Bogor Regency	9.89	10.54	10.94	11.8
3	Bogor Regency	32	23.08	23.14	24.62
4	Bogor Regency	5.95	7.78	6.87	7.97
5	Bogor Regency	6.07	8.37	6.41	5.05
6	Bogor Regency	3.26	7.99	6.85	3.48
7	Bogor Regency	8.83	13	11.12	17.39
8	Bogor Regency	127.17	37.71	41.81	35.44
9	Bogor Regency	3.43	7.16	4.86	3.54
10	Bogor Regency	7.15	12.56	10.67	10.09

Table 4. Low GRDP

No.	**Regencies/Cities**	**The average monthly household per capita expenditure (in IDR 100,000).**			
		Actual	GRF Prediction	RF Prediction	GLMM Prediction
Low GRDP					
1	Kuningan Regency	10.02	9.52	9.36	9.00
2	Kuningan Regency	9.47	11.79	12.48	12.99
3	Kuningan Regency	33.69	25.47	33.49	23.66
4	Kuningan Regency	6.43	8.65	7.89	8.16
5	Kuningan Regency	12.09	9.40	11.68	7.04
6	Kuningan Regency	7.44	9.44	10.16	8.42
7	Kuningan Regency	6.82	9.60	8.14	8.75
8	Kuningan Regency	8.59	9.23	8.58	6.54
9	Kuningan Regency	12.99	11.45	11.46	11.86
10	Kuningan Regency	10.11	9.94	10.54	8.41
1	Sukabumi City	25.16	30.05	33.07	33.89
2	Sukabumi City	5.83	15.18	17.36	15.20
3	Sukabumi City	23.46	14.11	17.17	15.38
4	Sukabumi City	12.58	9.51	10.06	9.33
5	Sukabumi City	3.53	7.90	5.62	6.65
6	Sukabumi City	6.60	8.44	7.09	10.14
7	Sukabumi City	5.18	14.25	10.96	16.93
8	Sukabumi City	4.69	11.39	9.21	15.55
9	Sukabumi City	29.93	23.56	28.79	20.38
10	Sukabumi City	10.40	22.17	18.68	24.22
1	Banjar City	7.59	11.64	11.57	12.43
2	Banjar City	7.34	14.35	12.01	13.68
3	Banjar City	6.41	12.18	9.85	13.90
4	Banjar City	17.44	15.00	12.39	19.29
5	Banjar City	7.23	11.38	8.53	12.79
6	Banjar City	6.39	8.97	8.09	8.76
7	Banjar City	22.70	21.28	22.44	22.66
8	Banjar City	9.89	10.18	10.41	9.96
9	Banjar City	9.26	9.88	10.26	12.73
10	Banjar City	13.75	11.85	12.74	14.80

Discussion

Based on the results of this study, the use of the Generalized Linear Mixed Model (GLMM) can be improved by considering other random effects, such as census block or sub-district. In this study, the random effect used was limited to the type of region (urban or rural), which may not capture more specific local variations. By incorporating random effects such as census block or sub-district level, it is hoped that future research using the GLMM model will provide better model performance and predictions.

CONCLUSIONS

The results of this study indicate that among the Generalized Random Forest (GRF), Random Forest (RF), and Generalized Linear Mixed Model (GLMM) methods, the best-performing method based on the smallest MSE and ANOVA testing is Random Forest (RF). RF outperforms GRF and GLMM in prediction accuracy, achieving the lowest MSE in both high and low GRDP regions. GRF ranks second, while GLMM has the weakest performance with the highest MSE, highlighting its limitations in handling household per capita expenditure data with random regional effects. RF provides the most accurate predictions with the smallest squared error, closely approximating actual values in most observed regions, making it the most reliable method for estimating per capita expenditure. The performance ranking remains consistent, with RF as the best method, followed by GRF, and lastly, GLMM, which exhibits greater errors. Additionally, household per capita expenditure estimates show that prediction errors are lower in the low GRDP category than in the high GRDP category. This further reinforces that RF is the most suitable method for modeling and estimating household expenditures in West Java.

REFERENCES

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[2] Chaudhary, K., Alam, M., Al-Rakhami, M. S., & Gumaei, A. (2021). Machine learning-based mathematical modeling for prediction of social media consumer behavior using big data analytics. *Journal of Big Data, 8*(1), 1-20. https://doi.org/10.1186/s40537-021-00301-0

[3] Conover, W. J. (1999). *Practical Nonparametric Statistics* (3rd ed.). Wiley. ISBN: 978-0471161104

[4] Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801131

[5] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics, 47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

[6] Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

[7] Goldman, N. (2022). Nonparametric Estimation of Conditional Densities by Generalized Random Forests. *Journal of Statistical Computation, 48*(3), 121–145. https://doi.org/10.1080/00949655.2022.2056264

[8] Zhang, Y., Li, H., & Ren, G. (2022). Estimating heterogeneous treatment effects in road safety analysis using generalized random forests. *Accident Analysis & Prevention, 165*, 106507. https://doi.org/10.1016/j.aap.2022.106507

[9] Wang, M., & Yang, Q. (2022). The heterogeneous treatment effect of low-carbon city pilot policy on stock return: A generalized random forests approach. *Finance Research Letters, 47*(Part A), 102808. https://doi.org/10.1016/j.frl.2022.102808

[10] Shiraishi, T. (2024). Time Series Quantile Regression Using Random Forests. *Machine Learning Journal, 113*(4), 789–805. https://doi.org/10.1007/s10994-024-06134-4

[11] Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Setiawan, D., Wijayanto, H., & Abdul Rahman, L. O. (2022). Bagging and random forest classification methods for unbalanced data school dropout cases in Lampung province. *AIP Conference Proceedings, 2662*(1), 020026. https://doi.org/10.1063/5.0111124

[13] Amaliah, S., Nusrang, M., & Aswi. (2022). Penerapan metode Random Forest untuk klasifikasi varian minuman kopi di kedai kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research, 4*(2), 121–127. https://doi.org/10.24815/variansi.v4i2.22865

[14] Ilma, H., Notodiputro, K. A., & Sartono, B. (2023). Association rules in random forest for the most interpretable model. *Barekeng: Journal of Mathematics and Its Applications, 17*(1), 185–196. https://doi.org/10.30598/barekengvol17iss1pp0185-0196

[15] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

[16] Rusyana, A., Notodiputro, K. A., & Sartono, B. (2021). A generalized linear mixed model for understanding determinant factors of student's interest in pursuing bachelor's degree at Universitas Syiah Kuala. *Jurnal Natural, 21*(2), 193–205. https://doi.org/10.24815/jn.v21i2.23818

[17] Sunandi, E., Notodiputro, K. A., & Sartono, B. (2022). A study of generalized linear mixed model for count data using hierarchical Bayes method. *Media Statistika, 14*(2), 194–205. https://doi.org/10.14710/medstat.14.2.194-205

[18] Belinda, N. S., Notodiputro, K. A., & Soleh, A. M. (2024). BHF and Copula Models in Small Area Estimation for Household Per Capita Expenditure in Bogor District. *Jurnal Natural, 24*(2). DOI: 10.24815/jn.v24i2.37278.

[19] Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example* (5th ed.). Wiley. ISBN: 978-1118841286

[20] Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal, 6*, 100164. https://doi.org/10.1016/j.dajour.2023.100164

[21] Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. *Joint Statistical Meetings, 12*(1), 3455–3460.

[22] Zubedi, F., Sartono, B., & Notodiputro, K. A. (2022). Implementation of winsorizing and random oversampling on data containing outliers and unbalanced data with the random forest classification method. *Jurnal Natural*, 22(2), 108–116. https://doi.org/10.24815/jn.v22i2.25499.

[23] Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). Sage Publications. ISBN: 978-1483381071

[24] Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd ed.). Academic Press. ISBN: 978-0123850719

[25] Field, A. (2013). *Discovering Statistics Using SPSS* (4th ed.). Sage Publications. ISBN: 978-1446249198